UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
REGION 10
OREGON OPERATIONS OFFICE
811 S.W. 6th Avenue
Portland, Oregon 97204

July 6, 2006

Mr. Jim McKenna
Port of Portland & Co-Chairman, Lower Willamette Group
121 NW Everett
Portland, Oregon 97209

Mr. Robert Wyatt
Northwest Natural & Co-Chairman, Lower Willamette Group
220 Northwest Second Avenue
Portland, Oregon 97209

Re: Portland Harbor Superfund Site; Administrative Order on Consent for Remedial Investigation and Feasibility Study; Docket No. CERCLA-10-2001-0240. Ecological Risk Assessment Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests

Dear Messrs. Wyatt and McKenna:

EPA has completed its review of the Ecological Risk Assessment Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests (Benthic Toxicity Interpretive Report). The document, prepared by Windward Environmental LLC for the Lower Willamette Group (LWG), is dated March 17, 2006. EPA comments are attached.

The recommended Sediment Quality Values proposed by LWG were based on the floating percentile method based on 3 of the 4 toxicity test endpoints (Hyalella growth was not included). NOAA, on EPA's behalf, developed alternative logistic regression models, using a larger freshwater database for the *Hyalella* 28-day growth and survival endpoint and calibrated these models to the Level 2 Effect Level in the Portland Harbor data. Both approaches were reasonably successful at developing a predictive relationship between sediment chemistry and toxicity. The two predictive modeling approaches were in agreement approximately 75% of the time and are useful for focusing in on areas where sediment contamination is likely to pose a risk to the benthic community.

In order to avoid schedule delays associated with production of the Round 2 Comprehensive Site Summary and Data Gaps Analysis Report (Round 2 Report), EPA recommends incorporating the results of the predictive model as presented with the following modifications:

1. Apply the alternative set of logistic regression models developed by NOAA on EPA's behalf to the Portland Harbor data set to improve the predictive ability of these tools.

2. Apply the approach recommended by the LWG (Floating Percenile Method) in conjunction with the alternative logistic regression models developed by NOAA as complimentary lines of evidence. Areas where both models predict risk or do not predict risk should be identified as such. Areas where the models are not in agreement should be identified as areas of indeterminate risk. Areas of indeterminate risk should be refined based on other lines of evidence used to evaluate risk to the benthic community.
3. The approach recommended by the LWG includes a proposed sediment quality values (SQV) of 1,270 mg/kg for total PAHs. This concentration is more than 50 times the concentration of the consensus based probable effects concentration (PEC) of 23 mg/kg developed by MacDonald and Ingersoll. As a result this value should not be applied to the data set. The LWG recommended floating percentile method should rely on the SQV developed for diesel range hydrocarbons as a surrogate for total PAHs.

The Round 2 Report should use the floating percentile methodology and the refined logistic regression methodology to identify areas of potential concern based on risks to the benthic community. Refinements to the predictive approach outlined in the attached comments should be used in conjunction with the results of the Round 2 report to identify additional data needs that will improve the models' ability to predict risks to the benthic community. These data gaps should be filled as part of the Round 3B sampling effort to be completed in 2007. EPA comments on the predictive models should be incorporated into the next iteration of the Benthic Toxicity Interpretive Report to be presented in the baseline ecological risk assessment and remedial investigation report. In addition, please submit a response to the attached comments within 60 days or contact us to discuss resolution of the comments.

Please contact Chip Humphrey at (503) 326-2678 or Eric Blischke (503) 326-4006 if you have any questions. All legal inquiries should be directed to Lori Cora at (206) 553-1115.

Sincerely,

Chip Humphrey
Eric Blischke
Remedial Project Managers

cc: Greg Ulirsch, ATSDR
Rob Neely, NOAA
Ted Buerger, US Fish and Wildlife Service
Preston Sleeger, Department of Interior
Jim Anderson, DEQ
Kurt Burkholder, Oregon DOJ
Rick Keppler, Oregon Department of Fish and Wildlife
Kathryn Toepel, Oregon Public Health Branch
Jeff Baker, Confederated Tribes of Grand Ronde
Tom Downey, Confederated Tribes of Siletz
Audie Huber, Confederated Tribes of Umatilla
Brian Cunninghame, Confederated Tribes of Warm Springs
Erin Madden, Nez Perce Tribe
Rose Longoria, Confederated Tribes of Yakama Nation
Valerie Lee, Environment International
Keith Pine, Integral Consulting

EPA Comments on Ecological Risk Assessment Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests
July 6, 2006

INTRODUCTION:

EPA would like to commend the LWG on the amount of effort that went into preparation of the Ecological Risk Assessment Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests (Benthic Interpretive Report). In general, EPA believes that LWG's proposed approach will serve as a useful tool in assessing risk and informing remedial decision making at the Portland Harbor site. EPA has developed detailed comments on the predictive models described in the report and recognizes that the project schedule does not allow time for the comments to be incorporated into the evaluation of benthic toxicity planned for the Round 2 Comprehensive Site Summary and Data Gaps Analysis Report (Round 2 Comprehensive Report). In order to avoid schedule delays, EPA recommends incorporating the results of the predictive model into the Round 2 Comprehensive Report as presented with the following modifications:

1. NOAA developed alternative logistic regression models, using a larger freshwater database for the *Hyalella* 28-day growth and survival endpoint and calibrated these models to the Level 2 Effect Level in the Portland Harbor data. EPA notes that both approaches were reasonably successful at developing a predictive relationship between sediment chemistry and toxicity; the two predictive modeling approaches were in agreement approximately 75% of the time and are useful for focusing in on areas where sediment contamination is likely to pose a risk to the benthic community. EPA believes that this alternative set of logistic regression models should be applied by the LWG to the Portland Harbor data set to improve the predictive ability of these tools.
2. The approach recommended by the LWG (Floating Percentile Method) should be applied in conjunction with the alternative logistic regression models developed by NOAA as complimentary lines of evidence. Areas where both models predict risk or do not predict risk should be identified as such. Areas where the models are not in agreement should be identified as areas of indeterminate risk. Areas of indeterminate risk should be refined based on other lines of evidence used to evaluate risk to the benthic community.
3. The approach recommended by the LWG includes a proposed sediment quality values (SQV) of 1,270 mg/kg for total PAHs. This concentration is more than 50 times the concentration of the consensus based probable effects concentration (PEC) of 23 mg/kg developed by MacDonald and Ingersoll. As a result this value should not be applied to the data set. The LWG recommended floating percentile method should rely on the SQV developed for diesel range hydrocarbons as a surrogate for total PAHs.

The Round 2 Report should use the floating percentile methodology and the refined logistic regression methodology to identify areas of potential concern based on risks to the benthic community. Refinements to the predictive approach outlined in the attached comments should be used in conjunction with the results of the Round 2 report to identify additional data needs that will improve the models' ability to predict risks to the benthic community. These data gaps should be filled as part of the Round 3B sampling effort to be completed in 2007. EPA

comments on the predictive models should be incorporated into the next iteration of the Benthic Toxicity Interpretive Report to be presented in the baseline ecological risk assessment and remedial investigation report.

GENERAL COMMENTS

Focus the Modeling Efforts: This report recommends focusing on the floating percentile method for future modeling efforts. As described above, the LWG and NOAA models are in agreement approximately 75% of the time. As a result, EPA believes that both models should be utilized as complimentary lines of evidence. Areas where both models predict risk or do not predict risk should be identified as such. Areas where the models are not in agreement should be identified as areas of indeterminate risk. Areas of indeterminate risk should be refined based on other lines of evidence including empirical estimates of benthic toxicity using bioassays; comparison of benthic tissue data (empirical measurements or modeled through application of BSAFs) to tissue TRVs; comparison to consensus, empirical and/or empirical based sediment quality guidelines (SQGs); and comparison of transition zone water data (empirical measurements or modeled through application of partitioning equations) to AWQC or literature values.

***Hyalella* growth and survival endpoint**: The Lower Willamette Group (LWG) proposes to disregard the results of the *Hyalella* growth and survival (pooled) endpoint. LWG supports this proposal based on "difference from other endpoints" and "no correlation with mortality endpoint." Yet these are precisely the reason that multiple test endpoints are required (because different test endpoints may show different sensitivities to different chemical mixtures). However, there was substantial agreement between the *Hyalella* and *Chironomus* pooled endpoints for samples that showed an extreme degree of toxicity (e.g., < 50% of control) in either test. The "lack of correlation to Chemicals of Concern" and the "effect of percent fines" may be more related to the different contaminant mixtures and gradients in the Portland Harbor study area. In a complex environment with multiple chemical mixtures and gradients with limited numbers of samples from any one area, a lack of correlation between a test endpoint and individual chemicals does not necessarily imply that toxicity is not related to chemical contamination. This is supported by the differences in chemicals that "set" the different models for the same sample (for example, the chemical with highest ratio of concentration to floating point value for a sample may be a phthalate, while the chemical with the highest probability of toxicity in logistic regression models may be ammonia or DDT for the *Hyalella* pooled model or PCBs or cadmium for the *Chironomus* pooled model). Because each contaminant can be considered as an indicator of toxicity for the chemical mixtures, it is not surprising that generic indicators such as percent fines, ammonia, or sulfides are good predictors of toxicity.

Proposed total PAH threshold values: The proposed Effects Level 2 and Effects Level 3 concentrations for total PAH, which represent AET values, are unreasonably high (1270 ppm DW) and significantly higher than other published values. For example, the proposed value exceeds the consensus-based freshwater PEC for Total PAH (22.8 ppm DW; MacDonald et al 2000) by more than a factor of 50. Of the samples exceeding the PEC value, 73% have a Level 2 response or greater in one or both of the pooled endpoints and 86% for samples with at least 25% fines. If we exclude the *Hyalella* growth endpoint, 62% of the samples exceeding the PEC have

Level 2 or greater response compared to 65% of the samples with diesel concentrations exceeding the proposed FPM value of 340 ppm. While diesel concentrations may be a slightly better predictor of toxicity than total PAH for this dataset, total PAH concentrations much lower than the proposed AET values are reliable predictors of toxicity. The proposed values for total PAH serve no useful purpose and should be discarded.

Inclusion of Appropriate Data in the Model: Data for which bioavailability is an issue should not be included in the predictive model. For example, high concentrations of PAHs may be detected in the sediments, but are bound up in a less bioavailable fraction such as pencil pitch. This issue was raised previously by EPA and its partners in the context of including Port of Portland Terminal 4 data in the analysis for this reason. Including these samples in the analysis can greatly skew the model results, because effect is not correlated with bioavailable fractions in the sediment. Based on our review of the report, the inclusion of GASCO effect / concentration data may skew the model results. This site has the potential to contain many different bound PAH contamination including pencil pitch. However, these samples were still included in the model analysis. This results in the inclusion of "no hits" with very high concentrations of PAHs. Looking at the highest no-hit concentrations, the top 6 samples are all in the vicinity of the GASCO site. Examples include G-264, 1,708,600 ppb total PAHs, G-301, 1,250,500 ppb, and G178, 470,060 ppb. Conditions off GASCO are confounded by the mixture in sediments of these less-bioavailable fractions such as pencil pitch and weathered tar pieces tar along with more fresh PAH and coal tar fractions that are more bioavailable and elicit effects. These two conditions may be teased out by a re-analysis of the sediment samples off the site. Conditions off GASCO can also lead to variance in the toxicity test results that are too high to detect anything but very large differences (low power), resulting in statistically indeterminate results. The GASCO site had the highest incidence of indeterminate samples at all effects levels (Figure 2-2). If these effects cannot be teased apart, we could simply omit samples off the GASCO site from the analysis. For this site, it may be clear that due to the variability in the forms of the contamination that we cannot accurately predict toxicity off this facility.

Three Tiered Framework: Based on the inherent reliability problems associated with development of a single SQV, EPA recommends calculating two screening values; a low screen below which a sample shouldn't be toxic and a high screen above which it should be toxic. Optimization should be possible at these two ends of the spectrum. As noted previously, the two models are generally in agreement in predicting very toxic samples and those that are clearly non-toxic samples. However, we don't agree on the classification of the samples that fall in between these classifications. The values that fall in between these two classifications would be classified as "indeterminate", and would require empirical toxicity testing or the use of additional lines of evidence. The LRM is well suited to this. It could also be done with the FPM (as was done for the DRAFT Washington Freshwater criteria).

Alternative Approaches For Subsets of PH Sediments: As stated in the March 18th work plan (Section 9.2), there are areas for which the predictive approach would not apply in Portland Harbor. This could include the physical form of the contaminant (as mentioned above), or the localized presence of contaminations over smaller spatial scales in the ISA (e.g. pesticides around RM 7). The work plan states models or other approaches would be developed for these areas. However, this was not included in the report. Also, areas where volatile chemicals were

detected in sediments and may be contributing to toxicity, but not evaluated in this report should be examined.

Level 1 Biological Effects Level: The report states "*it is recommended that Level 1 not be used to set SQVs for Portland Harbor because it is relatively unreliable in accurately predicting effects and well below the cleanup levels set at other regional Superfund sites.*" EPA agrees that Level 1 Biological Effects Level values should not be used as target cleanup levels. However, Level 1 values should not be discarded, as they represent concentrations associated with low level effects and provide useful information for defining areas of concern. The incidence of Level 1 or greater effects increases with increasing probability of toxicity.

Single-threshold evaluation of reliability: The report relies exclusively on a single-threshold evaluation of "reliability" of sediment quality guidelines. The conceptual model that a single value can accurately distinguish between "good" and "bad" samples, while perhaps desirable, is not consistent with most environmental data. EPA agrees that minimizing false negatives and false positives is an important goal, but concentration-response relationships are usually continuous and multiple thresholds may provide better separation of false positive and negative concentrations. For continuous models, such as the logistic regression model, an evaluation based on a single-threshold loses important information.

LRM model development: The logistic regression models were developed following the published approach developed by NOAA and EPA (Field et al. 1999; Field et al 2002; EPA 2005). The model development presented in the report did not address exclusion of chemical models that resulted in a high degree of false positives or adjustments to the screening approach to reduce the influence of a small number of non-toxic samples with very high chemical concentrations, which was particularly problematic for PAHs. The models were evaluated for reliability using the single threshold approach. Although this evaluation provides some useful information, reducing the evaluation to a single threshold does not take full advantage of the continuous concentration-response relationship.

NOAA developed alternative logistic regression models, using a larger freshwater database for the *Hyalella* 28-day growth and survival endpoint and calibrated these models to the Level 2 Effect Level in the Portland Harbor data. EPA believes that this alternative logistic regression model should be applied by the LWG to the Portland Harbor data set to improved the predictive ability of these tools.

Recommended FPM values: The recommended FPM values are based on 3 individual endpoints (*Chironomus* survival, *Chironomus* growth, and *Hyalella* survival), excluding results for the *Hyalella* growth endpoint and for the combined (pooled) growth and survival endpoints for both test species. The pooled results are important to consider, because growth and survival are not independent measures. (See previous discussion of the rationale for including the *Hyalella* growth and survival combined endpoint.)

Several of the recommended FPM values have the same concentration for Level 2 and Level 3 Effects. This indicates that these values are at the upper end of the concentration-response relationship and thus may be considered extreme effect concentrations.

PEC-quotient approach: The report did not evaluate the PEC-quotient (PEC-q) approach (Ingersoll et al 2001) – one of the major approaches to developing freshwater guidelines – which has been applied effectively in other Superfund remedial investigations (e.g., Calcasieu Estuary, Louisiana). A quick review of the data indicate that samples with mean PEC-q's greater than 1 show a Level 1 response or greater in at least one toxicity test endpoint in 87% of the samples and at least a Level 2 response in 77% of the samples. This suggests that the PEC-q approach may be useful in contributing to the identification of areas of concern. Evaluation of the Ingersoll PEC-q should be performed to determine if it is useful for the Portland Harbor remedial investigation.

Data Gaps: TPH was found to have good potential relationships with toxicity. However, because TPH was only analyzed at a limited number of stations, the model cannot assess this relationship (see page 21). As a result, additional TPH data may be required.

SPECIFIC COMMENTS:

Page 1, Section 1.0, Introduction: There are statements made here that state that the sediment toxicity testing and derivation of sediment quality values (SQVs) form the primary lines of evidence for the benthic community, and that other lines of evidence such as tissue residue concentrations and comparison to surface water and transition zone water concentrations would be secondary lines of evidence. This text should be revised, as the weights of different lines of evidence will be developed through the development of the weighting matrix.

Page 5, Section 2.0, Data Quality and Organization: The report states that "*petroleum data for 203 stations*" were available. How were the 146 stations with matching toxicity data for petroleum analysis selected?

Page 5, Section 2.1.2, Biological Effects Definitions: The report states that "*The biological effects levels used in the analyses are intended to correspond conceptually to "no effects level" (Level 1), "minor effects level" (Level 2), and "moderate effects level" (Level 3). As requested by EPA (EPA 2005a), the three levels were set at 90, 80, and 70% of the response observed in the control sediment, respectively*." The biological effect levels are mischaracterized. A more appropriate characterization would be "minor effects level" (Level 1), "moderate effects level" (Level 2), and "severe effects level" (Level 3).

Pages 5-6, Section 2.1.2, Biological Effects Definitions: Previous comments submitted by EPA have expressed concern about the selected alpha level for determining statistical significance. According to the work plan proceeding this report (*Estimating Risks to Benthic Organisms Using Sediment Toxicity Tests*, FINAL, dated March 18, 2005), an alpha level of 0.1 was to be used where it is found that test power of the dataset is low, according to ASTM guidelines (2003). Only an alpha level of 0.05 was used here. Since power is directly related to variance in the sample data, the variance in the analysis should be clearly reported and understood. To address concerns about the appropriateness of the statistical analysis to determine hits and no hits, it is recommended that the methodology outlined by Thursby et al., 1997 and Phillips et al., 2001 be followed. This approach more directly deals with issues that hinder appropriate statistical

comparisons to determine statistical difference. This protocol considers performance over a large number of comparisons. MSD values are calculated to determine a critical threshold for statistically significant sample toxicity. Significant toxicity threshold values (as a percentage of laboratory control values) are presented for each species and endpoint based on the data.

Data should be reported as indicated in Table 2 of Phillips et al, 2001, which clearly shows the sample and control response, the sample response as a % of the control, MSD threshold, significance of t-test, and whether it was identified as toxic, non-toxic or indeterminate. This will improve the transparency of the statistical analysis, and will address several concerns associated with interpreting toxicity test data. These include:

1) The identification of small differences that are statistical different from the control, which may increase the probability of making a type I error (identifying a sample as toxic when in reality it is not). This reporting should eliminate cases where statistical significance is assigned in individual cases because the among-replicate variability is small in a more transparent fashion. It will allow for a better understanding for where and how much this occurred.

2) Samples with large variance in the data (e.g. variance lies outside the 10^{th} and 90^{th} percentiles) should be reported. Declaring a sample non-toxic in this case would lead to a greater probability of making a Type II error (saying it is non-toxic when in reality toxicity exists). This will help in understanding where areas of large variance occurred (e.g. where differences from the control exceed 20 to 25%), and further action in those areas such as re-testing.

3) This method more accurately describes Beta error through a graphic representation of the statistical power (1-B). For example, power curves can be developed using the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the variance. Power curves can be superimposed with curves showing the probability of statistical difference created from the cumulative frequency of calculated MSDs.
   This approach should explicitly define MSD for the project in a non-arbitrary manner. A better understanding of the power curves relative to the data's variance aids in decisions regarding what difference from control is appropriate for determining statistical significance. Once a threshold for significance is determine, all of the test's data is included in the acceptability analysis for that test.

Page 6, Section 2.1.2, Statistical Difference Determinations: What analysis was used to determine statistical significance? The footnotes on Table 2-1 state the means of untransformed mortality or weight data was used in the definitions of effect levels. Were test and reference stations tested for normality? Were t-tests used?

Page 6, Section 2.1.2, Indeterminate Stations: EPA understands that there may be situations where low power is a problem because the variance may be too high in the test replicates to detect anything but vary large differences. Since the test responses were compared to control responses for the statistical evaluation, it is likely large variability in response came from the test sediment. The source of the variance should be reported here, because it could be do to

variations in bioavailability related to chemical form in the environment, or due to poor sediment homogenization prior to testing.

Page 6, Section 2.1.2, Biological Effects Definitions, Statistical Difference from Negative Control: For the floating percentile analysis, it would be still important to include a Level I effects level based on a statistical difference from negative control. Again, this may be more important for the floating percentile analysis (and AET derivation), especially since it is so reliant on the how we define no-hits, as apposed to hits (see page 7, second paragraph). Very small magnitude differences at the low end of the effects range may be very important for the development of the floating percentile model. The logistic regression model is not as sensitive to the omission of hits at the low range because it is the prevalence of toxic samples that primarily drive the curves, and the development of model relationships are not adversely affected by low power samples (Jay, correct me if I am wrong).

Page 7, Section 2.1.3, Use of Historical Toxicity Data: The objectives of the modeling effort are not just to improve model reliability as defined in the footnote on page 6 (correct predictions / total stations). The results of combining historical or regional data should be presented in how it changes the endpoints the government team are interested in optimizing; including % Predicted No Hit Efficiency.

Page 8, Section 2.2.1 – Data Quality: The report states that "*The exclusion of data with the N-qualifier primarily affected the pesticide data. Between 23 and 53% of the data for the following pesticides were excluded: aldrin, hexachlorocyclohexane (alpha-, beta-, and delta-), nonachlor (cis- and trans-), dieldrin, and methoxychlor. Between 35 and 67% of the summed data of DDD, DDE, DDT, total DDT, total chlordane, and total endosulfan were excluded.*" Considering that some of these contaminants are known to be of importance in the Lower Willamette, further evaluation of the exclusion of the aforementioned results should be performed. For example, what percentage of the excluded data had concentrations that exceeded the 25th percentile of the detected/included data? Would including these data affect the results? .

Page 8, Section 2.2.1, Data Quality: The text states that results with qualifier definitions listed in Table 2-3 were excluded. It looks like excluding samples with the "N" qualifiers excluded a lot of data (esp. pesticides). It should be confirmed that all PCB / DDT interferences in this dataset were properly re-analyzed according to previous EPA direction and the memo entitled "*EPA Region 10 Guidance for Data Deliverables from Laboratories Utilizing SW-846 Methods 8081 and 8082 from the Analyses of Pesticides and PCB Aroclors*". High detection limits, or elimination of "N" qualifiers that may represent interference problems can have a significant affect on the appropriateness of any model that attempts to correlate effects with sediment concentrations. This is particularly worrisome because the text states that this exclusion primarily affected the pesticide data, and that "between 35 and 67% of the summed data of DDD, DDE, DDT, total DDT, total chlordane, and total endosulfan were excluded" (in addition to between 23 to 53% aldrin, hexachlorocyclohexane, nonachlor, dieldrin, and methoxychlor).

Since the "N" qualifier is not an undetected value, it is unclear if this was an appropriate exclusion. It is also unclear why "N" qualifiers combined with "T" values were excluded. Is this the result of combining the results of two different analyses on one sample, where both of

them were an "N"? Or, was one sample an "N" and the other a "J"? J values certainly shouldn't be excluded, so if they were combined with an "N" as a result of another analysis method shouldn't the "J" estimate take priority? Generally, EPA recommends including the N, NJ, and NJT values for modeling purposes.

Page 9, Section 2.2.2, Data Organization and Reduction: The report states that "*The presence of non-toxic, naturally occurring crustal elements such as aluminum and selenium can confound the development of meaningful SQVs for the remainder of the analytes*." It is not clear why this should be the case. This may be an issue for FPM development, but LRMs are developed independently for each chemical and the crustal elements can be included or not in the development of the maximum probability model.

Page 9, Section 2.2.2, Data Organization and Reduction: For the FPM, aluminum and selenium should be added back into the model. The analysis shows that there is an association between aluminum and effects. I also wouldn't say that just because they are crustal elements that they are non-toxic, or that they cannot also be elevated anthropogenically. The ANOVA results for these chemicals need to be included in Table 5-2. If they are not associated with toxicity, they will drop out if appropriate.

Page 11, Section 2.3.3 and Table 2-4: For some chemicals where there were elevated detection limits, the exclusion of these chemicals in contributing to the sum could underestimate the total concentration. In general, when summing chemicals ½ the detection limit should be used for non-detected values. In addition, the report states that "*Individual dioxins and furans (replaced by TEQ).*" TEQs are based on tissue concentrations may not be meaningful in sediment without accounting for differences in bioaccumulation factors for individual PCB, dioxin, and furan congeners.

Pages 11 – 12, Section 2.2.3 – Chemical Summation: The report states that "*Using summations reduces covariance problems, and past side-by-side comparisons of other Oregon and Washington data sets have shown better reliability when summations are used*." Please provide reference(s) in support of this statement.

Page 12, Section 2.2.4 – Normalization: The report states that "*Normalization of non-polar organic compounds and metals could be applied in an attempt to improve the reliability of the predictive model(s). However, no actual advantage has been revealed in past side-by-side comparisons of other Oregon and Washington data sets, and the reliability of the non-normalized sediment quality guidelines is generally the same or better than the normalized guidelines*." Please provide reference(s) in support of these statements.

Page 12, Section 2.2.4, Normalization: It is unclear if normalization was tried for this dataset to determine if it improved the reliability estimates.

Page 13, Section 3.0 – Comparison to Existing Sediment Quality Values: Evaluation of the performance of paired values, such as TELs and PELs, using a single threshold is inappropriate. These types of sediment quality guidelines were developed to provide a lower level below which toxicity would be unlikely and a higher level above which toxicity would be likely.

Page 13, Quotient Methods: The report states that "*Quotient methods were developed as an approach to increase the predictive ability of certain SQVs (Long et al. 1998)*" Please refer to and cite the key papers on development and application of freshwater quotients (Ingersoll et al 2001; MacDonald et al 2000). NOAA suggests that it would be useful to apply the PEC-q method presented by Ingersoll and MacDonald to the Portland Harbor data.

Page 16, Section 3.2 – Reliability Analysis Results: The report states that "*In general, the quotient methods are an improvement over most of the SQV sets discussed above although not sufficiently reliable for use in predicting toxicity results at this site (see Appendix A). It is possible that the quotient approach has merit, but it needs to be optimized on a site-specific basis.*" A quick review of the data indicate that samples with mean PEC-q's greater than 1 show a Level 1 response or greater in at least one toxicity test endpoint in 87% of the samples and at least a Level 2 response in 77% of the samples. This suggests that the PEC-q approach may be useful in the identification of areas of concern. EPA requests that the LWG present the results of the PEC-q analysis conducted by LWG for this report.

Page 16, Section 3.2 and Appendix A, Section A.3.4., Quotient Methodology: The quotient method should be more thoroughly explored with the Portland Harbor data. Based on the results here, it seems to have merit, although the methodology and analysis here was not fully explored. This analysis is important because it quotients are designed to analyze chemical mixtures.

Page 17, Section 4.0, Exploratory Analyses to Support Development of Site-Specific SQVs: Please explain what is meant by the term "*chemical endpoints*"?

Page 18, Section 4.1, Last Paragraph: Why was magnitude of toxicity only evaluated for these chemicals that had greater then 50% detection frequency?

Page 18, Section 4.1 – Statistical Correlations: The report states that "*Even if correlations were not highly linear throughout the range, it was true for nearly all chemicals that high concentrations occurred in sediments with the highest fine-grained fractions (i.e., high concentrations implied high percent fines, but high percent fines did not always imply high concentrations).*" This also implies that, in general, high percent fines are a good indicator of high chemistry and that low percent fines are good indicator of low chemistry.

Figure 4-1: Figure 4-1 is not clearly explained. For example, it is unclear whether everything was correlated with everything in the table and only the highest correlations identified.

Page 22 and 23, Section 5.1, Floating Percentile Methodology: Although the Portland Harbor specific FPM reliability was compared to the reliability of other SQGs (e.g. TECs, PECs) and regional numbers generated by Washington State (e.g. SQS and CSL) in Appendix A, what is missing is an analysis and reliability of the combined datasets. The Portland Harbor data consists of about 220 bioassay results. Given the variability in the Portland Harbor system, 220 samples may not entirely represent the range of contaminants and conditions sufficiently to develop a Portland Harbor specific model. However, combining this dataset with other relevant regional data could help better refine the model and help fill in areas of the Portland Harbor data

that represent a limited range of concentration and toxicity. The fact that some SQVs stay the same between effect levels (e.g. level 2 and 3) may indicate that only a small subset of the range was tested here, and not solely due to the concentration-response curves (see page 65, second paragraph). This analysis may be especially important for the floating percentile methodology since the calculation of SQVs using this method is so dependent on the characteristics of the dataset under evaluation.

Page 23, Section 5.1 – Floating Percentile Model: The report states that "*These ranges may overlap due to site-specific or sample-specific variations in bioavailability or toxicity.*" This statement appears to assume causality, which may not be the case. The concentrations for a chemical that are associated with toxicity may have at least as much to do with the mixtures of other chemicals present in the sample as bioavailability. The report further states "*...and this is the source of most of the false positive errors.*" This statement requires clarification.

The report also states in this section "*Above the red bar, both false negatives and false positives may occur, as is shown for Chemicals A, B, and C. This region is the range of concentrations over which sample-specific bioavailability plays an important role in toxicity,...*" Please explain the basis for the bioavailability assertion. Does this assume causality for individual chemical concentrations?

Page 24, Section 5.1.1 - FPM Methodology: The report states that "*...hand-optimization steps were used to identify chemical concentrations for each endpoint and effects level in order to minimize prediction errors.*" Please explain further how this was accomplished?

Page 24, Section 5.1, Floating Percentile Model - Step 4: The analysis presented here, consistent with earlier FPM applications, splits the hit and no-hit groupings into different distributions before calculating the percentages. This does not seem like a necessary or justifiable approach. In applying the floating percentile method, it makes sense to include the larger entire dataset. Limiting the distribution to the smaller no-hit dataset would appear to reduce the ability of the method to refine the resulting optimized concentrations. EPA recommends that a single distribution of all the data be used in calculating the percentages. The use of one distribution to develop screening values does change the SQVs and the reliability estimates. The false negative and predicted no-hit efficiency values are similar to the LWG results, but in some cases the false positive and predicted hit efficiency values are substantially different (higher). These concerns were raised early in the evaluation of this model with the LWG, but the report fails to comment or acknowledge these concerns.

In addition, how were the results for each station assigned a hit/no hit status to create the distributions? Was this based on statistical difference only?

Step 5 states that analytes were only retained for model development for each endpoint if they were associated with toxicity at two or three of the effects levels. Why was this limitation placed on the analysis? It seems like we would want to pursue associations of toxicity even if it was only at one effects level. For example, these criteria exclude the development of models for cadmium, diesel-range hydrocarbons, mercury, pentachlorophenol and total PAHs for the *Hyalella* growth endpoint.

If the distributions were determined not be statistically different, then the contaminants were assigned AET values. It seems like they may not be statistically different because of the variance in response, but it may still not be appropriate to select the highest no effect concentration as an AET. If this is due to variance, it may be more appropriate to rely on empirical tests to determine toxicity at a given location.

Step 6: All hand optimizations need to be documented.

Page 26, Section 5.1.1 - FPM Methodology: The report states that "*Certain chemicals had no significant differences for any of the hit/no-hit definitions or endpoints. These included: 4-methylphenol, aldrin, alpha- hexachlorocyclohexane, antimony, bis(2-ethylhexyl)phthalate, butylbenzyl phthalate, chromium, delta-hexachlorocyclohexane, dibutyltin, hexachlorobenzene, monobutyltin, pentachlorophenol, phenol, tetrabutyltin, total dioxins/furans, total endosulfans, and tributyltin.*" It appears that this statement is not consistent with the results in Table 5-2 for at least 4-methylphenol, antimony, and pentachlorophenol. Please check and revise accordingly or provide clarification.

Page 29, Section 5.1.1 - FPM Methodology: The report states that "*It is also interesting to note that for most endpoints, bulk petroleum (diesel-range hydrocarbons and residual-range hydrocarbons) was somewhat more strongly correlated with toxicity than were total PAHs, in spite of the fact that PAHs were measured at all stations, and bulk petroleum was measured at only a subset of stations.*" Diesel- and residual-range hydrocarbons were only measured at selected stations. What was the basis for selecting the stations for the petroleum hydrocarbon analysis? For the stations selected for hydrocarbon analysis, diesel and total PAH were strongly correlated. [The average total PAH concentration was much higher for samples with diesel measured, approximately 126 ppm compared to 2.6 ppm for the other samples.]

Page 30-31, Section 5.1.1: EPA has been working with the Oregon Department of Environmental Quality (DEQ) to reproduce the FPM. In addition to using different distributions, as mentioned previously, there are a couple of other things that the DEQ FPM model handles differently. We have not had time to make the DEQ version of the model the same, but these could also be contributing to some of the differences that we have been seeing.

- When increasing the concentrations in an attempt to lower the number of false positives, we took the following steps : (1) find the chemical with the highest number of FPs; in case of a tie, use the chemical with the lower concentration in that step, (2) increase the concentration of that one chemical by the designated increment, (3) recalculate the %FN and #FP and (4) if %FN goes up, go back to previous step and consider that chemical completed, otherwise start over with step (1) until #FP reaches 0. From the description in this report, it appears that the LWG goes back to step (2) instead of (1). In other words, once the chemical with the highest #FP is selected, they keep raising the increments and recalculating the %FN and #FP until that can no longer be done for that chemical. At that point, they select the next highest #FP. This difference could result in our getting different values.

- **Page 31, 5th Bullet:** When increasing the %FN to the next level, they build on the values determined in the previous step. In the DEQ model, each step is independent. After 5% FN, it just starts over at 10% FN without any regard to the previous answers. DEQ was definitely not aware of this difference before submittal of this report, and this explains how they avoid getting answers that sometimes go up and down instead of going up or holding steady when[?] the %FN is increased. In previous versions of the FPM, there was up and down variability in the answers. This is a new step to prevent it from occurring with this dataset. However, artificially determining to build on values from the previous step seems arbitrary and may hide problem areas with the methodology.

- DEQ's main concern with the differences in our results is that fact that we do not get the same results for our performance measures and our results are not as high as reported here. This appears to be related to the fact that DEQ ends up with higher numbers of FPs, thus creating lower values for Efficiency (No-Hit Reliability) and Predicted Hit Reliability. This may also be tied in with the slight differences in steps mentioned above.

Page 36, Section 5.1.2, Results of FPM Runs: The report states that "*...there are a limited number of analytes for which FPM values can be calculated because the level at which these analytes reach their toxicity threshold is apparently above their concentration ranges in this data set.*" The term "toxicity threshold" appears to assume causality for an individual chemical. In environmental mixtures, this is an unjustified assumption.

Page 43, Section 5.3, Logistic Regression Analysis: It does not appear the LWG performed a separate optimization (curve fitting) of the Pr_Max values with proportion of toxic samples. In addition, Pr_Prod has been dropped in previous evaluations as a reliable measure, but we would like to see an evaluation in LWG's report. On basic principles, we would expect to see this work well. Perhaps it will end up working better with the Portland Harbor dataset.

Page 43, Section 5.3.1: The "screened data set" method mentioned in Step 4 of the LRM Methodology should be evaluated to see if it would have any beneficial results for the FPM.

Page 53, Section 5.3.2 – Results of the LRM Runs: The report states that selection of a single threshold from a continuous relationship is not a useful application of these models.

Page 55, Section 5.3.2 – Results of the LRM Runs: Regarding the reference to "*Chemical drivers*", please clarify that "chemical drivers" refers only to chemicals that play a role in the predictive model (i.e., the best predictors of toxicity of the chemical mixtures in the study area) and may have nothing to do with "chemical drivers" of toxicity.

Page 56, Section 5.3.2 – Results of the LRM Runs – Influence of Grain Size: The report states that "*An effect of grain size on toxicity is seen only for Hyalella pooled at Levels 2 and 3. This correlation between the Hyalella pooled and percent fines is indicated by the presence of percent fines as a chemical driver.*" A correlation with percent fines does not demonstrate a grain size effect and does not imply that percent fines is causing toxicity. The highest concentrations for each chemical are associated with samples with high percent fines, so it

cannot be concluded that fines are causing toxicity in the *Hyalella* pooled endpoint. (See next comment about the use of the term "chemical drivers".)

Page 56, Section 5.4 – Discussion of Chemical Drivers: Regarding the reference to "Chemical Drivers": Chemicals that are good predictors in the models should not be assumed to be causing toxicity. The report should make a clear distinction between chemicals that are "drivers" in the models and those that are associated with causality. Please revise accordingly.

Page 57, Section 5.4 – Discussion of Chemical Drivers: The report states that "*Ammonia and sulfides are common confounding factors in bioassays (ASTM 2003) and can sometimes be high enough to cause toxicity in bulk sediments, even when their levels in overlying water are below bioassay QA/QC criteria*." Please clarify the basis for the statement in the 2nd part of this sentence. Does information exist which shows that the bioassay QA/QC criteria values for ammonia and sulfides in overlying water are too high?

Pages 58 - 59, Section 6.1 – Methods not Retained for Use: The report states that "*...it became clear that the Hyalella growth endpoint was responding differently than the other endpoints from a variety of standpoints, which raised some concerns*." Isn't this a primary reason for using different toxicity endpoints?

Page 59, Section 6.1 – Methods not Retained for Use – Effect of Percent Fines: A correlation does not demonstrate an effect. As pointed out earlier, most of the high chemistry was found in high percent fines samples. Please change "effect of" to "correlation with" or similar term that does not imply causality. The report also includes the following statement: "*Certainly, there are precedents for high- and low-percent fines effects on other amphipods, both freshwater and marine, in commonly used toxicity tests*." Please provide reference sources for this statement.

Page 60, Section 6.1 – Methods not Retained for Use - Level 1 Biological Effects Level: The report states that "*The reliability of nearly all the endpoints at Level 1 is reduced as compared to Levels 2 and 3. This is likely due to the very small difference (10%) from control used to define the Level 1 endpoints. This level of difference is likely within natural and laboratory variability in many cases*"...A difference of 10-20% from control was statistically determinate for most of the samples for all endpoints, indicating that it was outside the range for laboratory variability for the tests conducted. The Level 1 Biological Effects Level is useful for identifying concentrations at the lower end of the concentration-response relationship, in contrast to the Level 3 concentrations, which are at the upper end of this relationship.

Page 62-63, Section 6.3 – Floating Percentile Model - Sensitivity to individual chemicals varies by endpoint: The report states that "*The chemicals that showed a relationship to toxicity varied by endpoint. The Chironomus growth, Chironomus mortality, and Hyalella mortality endpoints were sensitive to similar chemicals, while the Hyalella growth endpoint showed a very different relationship*." Individual chemical sensitivity should not be asserted or implied from correlations with environmental chemical mixtures. EPA suggests using terminology that refers to the relationship between toxicity endpoints and chemical concentrations as "correlation" or "association."

Page 63, Section 6.3 – Floating Percentile Model: The report states that "*The results of this model correspond well both with measured toxicity and with the conceptual site model.*" NOAA is not clear on the meaning of this statement. In what way or how does the model correspond well with measured toxicity and the conceptual site model? Does this mean the model corresponds well with measured toxicity and those locations where one would expect to see toxicity based on the conceptual site model? Which conceptual site model(s) (ecological, human health, overarching CSM)? Please clarify.

Page 64, Section 6.4 – Proposed Sediment Quality Values: This section should distinguish between contaminants that were included in the model, but were not good predictors of toxicity, and those that were not included in the model because there were less than 30 detections. Those with less than 30 detections may still be of interest in toxicity identification, just on a smaller scale.

Page 65, Section 6.4 – Proposed Sediment Quality Values: The report states that "*Bulk petroleum measures were more strongly correlated with toxicity than total PAHs, even though PAHs were measured at all stations, and bulk petroleum was measured at only a subset of stations. Although the SQVs for PAHs may appear high, they are consistent with those derived from other West Coast data sets (e.g., San Francisco Harbor (Germano & Associates 2004), Los Angeles Harbor (unpublished)) using the FPM and the LRM, indicating that PAHs alone are not large contributors of toxicity to benthic organisms. PAHs are only a small subset of the suite of narcotic chemicals present in sediments and in petroleum, all of which may affect benthic organisms through similar toxicological pathways (McCarty 1991; McCarty and Mackay 1993; McCarty et al. 1992). The bulk measures of petroleum appear to better capture and correlate with that toxicity, as is apparent from the SQVs calculated for these measures.*" In Los Angeles Harbor as well as the entire California Sediment Quality Objectives database, the total PAH concentrations were much lower – very few samples exceeded the ERM of 44 ppm and none were within an order of magnitude of the proposed values. The LRM results for Los Angeles Harbor showed that PAHs infrequently had the maximum probability for a sample, but the logistic regression model probability of toxicity associated with the proposed PAH SQV would be very close to 1 (maximum possible). NOAA is concerned that the statement, as presented, is inaccurate and/or incorrect. NOAA is adamant that the presented SQVs for PAHs are not acceptable.

The report states that "*The FPM often identifies similar values for different effects levels, as can be seen in Table 6-1 (this is also true of AETs). Some chemicals, such as ammonia, arsenic, and residual-range hydrocarbons, have different SQVs at Level 2 and Level 3. Other chemicals, such as copper, diesel-range hydrocarbons, and DDTs, have the same SQV at both levels. Although at first this may appear unusual, it reflects the fact that the concentration-toxicity curve for these chemicals is apparently steep in Portland Harbor.*" Please provide the factual basis for this statement? Consider that this result may be interpreted to suggest that the similar values for different effects are near the upper end of the concentration-response relationship. This is certainly the case for total PAH.

Appendix A: Appendix A states: *"For each existing SQV set, the more protective of the two thresholds (TEL, TEC, LEL, and SQS) was compared to the Level 1 and 2 biological effects levels, and the higher of the two thresholds (PEL, PEC, SEL, and CSL) was compared to the Level 3 biological effects levels, consistent with the narrative intent of these SQVs." The PEL and PEC SQGs should be compared to all three biological effect levels to be consistent with the data used in their derivation and their narrative intent*." Consistency with the narrative intent for paired guidelines would preclude calculating reliability based on a single threshold. The TEL-type thresholds should be evaluated for their reliability in predicting the lack of toxicity and the PEL-type thresholds for their reliability in predicting toxicity.

REFERENCES

Field LJ, MacDonald DD, Norton SB, Ingersoll CG, Severn CG, Smorong D, Lindskoog R. 2002. Predicting amphipod toxicity from sediment chemistry using logistic regression models. Environ Toxicol Chem 21(9): 1993-2005.

Field LJ, MacDonald DD, Norton SB, Severn CG, Ingersoll CG. 1999. Evaluating sediment chemistry and toxicity data using logistic regression modeling. Environ Toxicol Chem 18:1311-1322.

Ingersoll, C. G., D. D. MacDonald, et al. (2001). "Predictions of sediment toxicity using consensus-based freshwater sediment quality guidelines." Archives of Environmental Contamination and Toxicology **41**(1): 8-21.

MacDonald, D. D., C. G. Ingersoll, et al. (2000). "Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems." Archives of Environmental Contamination and Toxicology **39**(1): 20-31.

U.S. EPA (2005). Predicting toxicity to amphipods from sediment chemistry. National Center for Environmental Assessment, Washington, DC; EPA/600/R-04/030.